DOCUMENT RESUME

ED 209 762 EA 014 225

AUTHOR	Coffman, William E.
TITLE	Those Achievement Tests--How Useful?
INSTITUTION	Iowa Univ., Iowa City. Inst. for School Executives.
PUB DATE	80
NOTE	5p.; Address delivered at a Testing Conference sponsored by the Midcontinent Regional Educational Laboratory and the Regional Offices of Educational Programs VII and VIII (Kansas City, MO, May 1979).
AVAILABLE FROM	Institute for School Executives, University of Iowa, 210 Lindquist Center, Iowa City, IA 52242 ($1.00).
JOURNAL CIT	Executive Review; v1 n1 1980
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Academic Ability; *Achievement Tests; Criterion Referenced Tests; *Educational Testing; Elementary Secondary Education; Minimum Competency Testing; Norm Referenced Tests; *Standardized Tests; *Testing Problems; Test Interpretation; Test Reliability; Test Theory; *Test Use
IDENTIFIERS	School Effectiveness

ABSTRACT

Standardized achievement tests are often misused as indicators of a school's quality or effectiveness relative to other schools. This is an incorrect use because it ignores variation among schools in student abilities, family support of education, student mobility, and other factors. People also misuse tests because they impute to them more statistical reliability and validity than they actually have. An example using the Iowa Tests of Basic Skills illustrates the possibilities for reliability errors with both single tests and retesting over time. The use of norm-referenced tests to evaluate Title I programs, under the Elementary and Secondary Education Act, is incorrect because it makes the doubtful assumptions that national norms are valid and that one can compare either different groups of students or students with and without Title I treatment. However, criterion-referenced tests have similar problems--with reliability, neglect of some learning areas and processes, and arbitrary choice of satisfactory performance levels. Hence it is unsound to use criterion-referenced tests alone in minimum competency testing, because the tests do not account for the range of student abilities, the student's employability, or educators' possible failure to solve a student's learning disability. (Author/RW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Volume I, Number 1

The Executive Review

Institute for School Executives — The University of Iowa

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

W. Bozeman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED209762

EA 014 225

THOSE ACHIEVEMENT TESTS —HOW USEFUL?*

William E. Coffman

It's a good idea for a college professor to keep in touch with things beyond the campus, especially if he's involved with service activities like the Iowa Testing Programs. Not long ago, I had this brought sharply to my attention when I ran into an old acquaintance, Charlie Urbanfringe.

"Hi, Professor," he greeted me in his usual open and enthusiastic manner. "How are things going with the Iowa Testing Programs?"

"O.K., I guess," I responded. "I'm not sure just how long that will be the case, however," I added. "Did you see in the papers about the ISEA passing a resolution calling for a moratorium on standardized testing?"

"What's the ISE..."

"Well, it's the big organization teachers in this state belong to," I explained, "and apparently the leaders think the way some people are using tests in the schools is all wrong and we ought to stop using the tests till people learn how to use them right."

"That's funny," Charlie replied, a puzzled look on his face. "How in the world can they really tell how good they are if they don't use standardized tests?"

"Take our own school system here in Exurbia," he continued. "Our superintendent always calls a news conference as soon as the test results come back from the scoring center, and I must say, it makes me feel very reassured when the papers report that the schools in Exurbia are at the 95th percentile of all the schools in the state. The schools here certainly must be doing something right to stay up there near the top year after year."

"I saw where a committee of parents in Port City was calling for the resignation of their superintendent," he continued. "It seems that the Port City test scores just keep getting worse all the time, and the parents can't understand why they're so far below the norm. I heard that there was some talk of not reporting the scores for Port City to the newspapers at all, but the attorney general's office ruled that the summary statistics based on the testing program was public information."

"Now look, Charlie," I remonstrated, "test scores aren't as simple as they seem."

"Is that so?" said Charlie, and I could see that he wasn't taking my remark very seriously. I didn't have long to see why, for Charlie pulled me by the arm and I could see that he was going to report on one of his periodic brainstorms. "I've been thinking that nobody ever did a better job of making things simple than you educational testers. You report both percentiles and grade scores for each pupil so that a parent can see immediately just where his child stands in the group and just how many months ahead or behind he is each time he's tested. What could be better?"

"I was just thinking the other day how nice it would be if we could have the same kind of information on the doctors here in Exurbia—and in Port City, too, for that matter. You know, the doctors are always giving tests to their patients. Now suppose that instead of having the test results reported in numbers that only the doctors can understand, they had to report them in percentiles. We could then find out how the doctors stacked up. Boy! Give me the doctor with a percentile rank of 95. He must be doing something right!"

"Wait a minute!" I shouted, trying to interrupt the brainstorm but to no avail. He was wound up and just had to get his thoughts out.

"The only thing is," he continued, "I don't see how I could find out just how much better the 95th percentile doctor is. I haven't yet figured out how to get something like the grade scores you fellows have worked out for the schools."

At that point he stopped, and I had a chance to get a word in. "Look," I said, grabbing him by the lapels of his coat, "didn't it ever occur to you that the best doctor might have the sickest patients?"

"Huh?" he responded, and I could see I had made a point. "No-o-o—, I guess not . . . but then there wouldn't be much relationship between the measures the doctor made on his patients and how good a doctor he was, would there?"

I could see the wheels were running in his head, so I didn't say anything more. Charlie has a lot of common sense, and once he's thinking along the right track, you can usually depend on him to work things out in a sensible way.

After what seemed like half-an-hour but was probably only two or three minutes, Charlie started in again. "But if you can't tell how good a doctor is by how healthy his patients are, how can you tell how good a school is by how much the pupils know? Kids learn a lot of what they know outside the school. And some kids learn faster than others. I'll bet most of our kids here in Exurbia have parents who help them with their homework and I'll bet many kids in Port City don't get much help at home. What good, then, is it to know how Exurbia stands in relation to Port City on average test scores?"

"Well," I said, "it doesn't help much to make that kind of comparison. It does help, however, to know whether or not your schools are getting worse or better from year to year."

"I'm not sure," said Charlie, another light flashing. "Take Port City, for example. Over the past ten years there have been a lot of people moving out of the city and others moving in. I bet that's why the test scores keep going down."

"You have a point there," I said, making a mental note that I'd better not sell Charlie short.

"I suppose you could compare Port City with other cities like it, couldn't you?" Charlie continued, "but how many cities are just like Port City? And then you'd never be able to tell whether the schools of a community like Exurbia were better than the schools of a place like Port City, could you? Maybe the ISEA is right. What good are the tests if you can't use them to tell how good a job the schools are doing?"

"Well, in the first place, most of the standardized tests in use today were made to help teachers tell how much the pupils are learning as

*An address delivered at a Testing Conference sponsored by the Midcontinent Regional Educational Laboratory and the Regional Offices of Educational Programs VII and VIII, Kansas City, Missouri, May 7, 1979.

individuals, not to measure directly how good a job a school is doing," I explained.

"Oh, I see." Charlie was glowing again but not for long. Puzzled, he spoke softly as if to himself. "If a kid stays at the same percentile year after year he must be holding his own in the race. I guess that means that if he has the same percentile rank this year as he had last year, he's increased his score one grade since last year, but how far behind has he dropped if he loses 5 percentile points, or how much more has he grown if he gains 8 percentile points? I guess I could find that out by looking at his grade scores?"

"Wait a minute," I cautioned. "It isn't really that simple. In fact, I think educational tests are probably as complex in their way as the tests the doctors use. Sometimes I wish they didn't seem so simple. Then maybe people wouldn't think they understand them when they don't."

"But then, that means that people who use test scores need to know a lot more than I know, doesn't it? . . . Come to think of it, I don't usually ask my doctor what *scores* I got on the tests he uses. Sometimes he *does* tell me just so I'll know in case I have to see another doctor sometime. I remember he did tell me that my blood pressure was 160 over 85 and that wasn't too bad for a man my age, but actually the numbers themselves don't mean much. I wonder if it would help if he told me that I'm at the 63rd percentile in blood pressure? No, I guess I'd rather just let the doctor tell me in words what the test says about my health."

"Don't you think you'd better let the teacher tell you what your child's test scores mean, too?" I suggested.

All at once Charlie stopped and took out his watch. "My gosh!" he exclaimed. "I'm due at a meeting in three minutes. Nice talking to you." And he disappeared into the crowd of shoppers.

Now I'm sure that some of you are thinking that I've been putting words into Charlie's mouth; and of course I have been. But I don't think I've been misrepresenting the situation. The public, including legislators as well as parents, are very much concerned about how well their schools are doing; and they think it should be possible to give a simple answer to the question and that standardized tests, *because* they are "standardized," ought to provide *that* simple answer. Unfortunately, as Charlie Urbanfringe learned, interpreting test results isn't as simple a job as it seems at first glance. Faced with the choice between having test results that are misused and having no test results, knowledgeable teachers will generally vote for no test results—as the resolutions calling for a moratorium on testing show. There is danger that public pressure for publication of test results may destroy the very information teachers and administrators need to make wise decisions.

Yes, as Charlie Urbanfringe learned, standardized tests are not as simple as they seem. They can be misinterpreted and misused—not only by laymen like Charlie, but also by teachers and administrators. Let's spend a few minutes considering in somewhat greater detail just how this comes about.

One of the most serious errors people make is to assume that tests, because they are standardized and come with impressive manuals crawling with numbers, and because they carry names that appear to focus on desirable learning—such names as vocabulary, reading comprehension, language skills, arithmetic problem solving, etc.—I say, because they come in such attractive and impressive packages, people assume that they are more accurate and relevant measures than they actually are. In the technical terms of the measurement specialist, people assume that the tests are more reliable and more valid than they really are. Let me give you some examples.

The Iowa Tests of Basic Skills are tests that compare favorably with any general tests of achievement used in elementary schools. The reading comprehension subtest is reported to have a reliability of around .85 when used at a particular grade level.[1] Each year scores are reported on this subtest for children all over the country. They are reported in percentile ranks or in grade equivalent scores such as "Your score was 57 (fifth grade, seventh month), which places you at the 40th percentile of sixth graders like yourself all over the country. The manual reports that the standard error of measurement for this test, in grade six for scores between 50 and 59, is 6.2 GE units. In other words, for one out of three pupils, the reported score will differ by more than 6.2 units either way from the true score one would obtain from averaging many different tests just like this but with different questions. That means that the true score for an individual with a reported score of 57 might very well be as low as 50 (a percentile rank of 24) or as high as 64 (a percentile rank of 57). In other words, the pupil with a score of 57 might very well actually be in the lower quarter of the class or well above average. It takes considerable experience to be able to treat test scores as indicators, as useful information in the context of other information one might have about a pupil. Few of us ever learn as much as we'd like to know about interpreting test scores; but most of us with any knowhow at all feel more comfortable with periodic test reports than if we have to depend on our subjective impressions entirely, particularly if we can have the results of other tests in a comprehensive battery as additional checks. For example, it helps to know about a child's vocabulary score when we're interpreting a reading comprehension score. The two scores together tell us more than either one separately.

Let's now consider the language subtests of the ITBS. As some of you may know, there are four subtests—spelling, capitalization, punctuation, and usage—and the composite based on all four has a reported reliability considerably higher than the reading comprehension subtest, between .90 and .94 for the typical grade with estimates of standard error of measurement ranging between 2.0 and 5.6. Thus one can obtain a relatively more accurate estimate of the ability of a pupil to answer multiple-choice questions about the mechanics of English composition than about the same pupil's ability to answer questions about passages read. But how much can one say about a pupil's ability to express thoughts in writing on the basis of scores on the language subtests? Let me just leave with you one thought. If teachers are working very hard to help pupils write down their thoughts and to revise and rewrite until the compositions are clear and correct, I'm reasonably happy using the language test scores as indicators. But if teachers are spending all the time available in drilling children to answer multiple-choice questions on spelling, punctuation, capitalization, and usage in order to increase test scores, I don't have much confidence in the test scores as indicators of writing ability. Unfortunately, with newspapers publishing test results with the implications that they reflect the quality of education in the community, it's very difficult for teachers and school administrators to avoid placing the teaching emphasis in the wrong places.

The problem of reliability is much less when we're looking at average scores for a class or a school rather than at scores for individuals—and that's one reason why some people feel more comfortable dealing with such averages. But again, as Charlie Urbanfringe learned, by focusing on output only and by assuming that the test is a comprehensive and valid measure of whatever it is named, test results can be misused. Often the school with below average test scores is doing a top-quality job while a school with above average scores is doing a mediocre job.

When attention is called to this problem there's a tendency to say, "Well, let's measure at the beginning of the year and again at the end and see which schools have the greater gains." Unfortunately, it's not that simple. Recall that we noted considerable error in any standardized test score. When two scores are involved as in testing at the beginning and at the end of a period of instruction, the gain scores end up with the errors of both measurements and so become even less dependable than the test scores themselves.[2] The problem is particularly serious if pupils are chosen for special treatment on the basis of the first test results. The fact is that if you take the half of a class that scores below average on the first test and give a second test the next day, you're almost certain to find that the scores have gone up on the average. If you wait till the end of the term for the second test, you're quite likely to think all the gains were the results of your special instructional efforts.

It was recognition of problems such as this which led the U.S. Office of Education to develop special guidelines for the evaluation of Title I programs in the schools of the United States. Unfortunately, the guidelines as finally approved and disseminated, are strong on statistical manipulation and weak on considerations of validity; and this weakness is likely to be overlooked because there's a tendency for all of us to stand in awe of complex statistical analyses and to give too little attention to the assumptions necessary for the analyses to provide true answers to our questions.

The guidelines recommend three evaluation models for Title I evaluations, each of which rests on assumptions, and these assumptions are not likely to be met in the situations in which the models are applied. For example, the most commonly used model assumes that the national norms for a test are appropriate references against which to compare the results of a Title I program. Is this, in fact, the case?

National norms are developed by pooling the test scores from samples of pupils from schools all over the country—schools that teach subjects using different textbooks, different sequences, and different emphases. When patterns of growth for individual schools are charted,

the patterns may be quite different from the one presented by the averages. Spurts and slowdowns in the patterns appear, one school showing larger than typical changes from grade 4 to grade 5 and smaller than typical changes from grade 5 to grade 6 while another school may show just the opposite. Generally, when teachers in a school are asked to consider the patterns in relation to the content of the test and the characteristics of their curriculum, they can account for the departures from the norms. Sometimes, however, they can't, and it is in these cases that the test results may be most useful in directing attention to relationships between curricular patterns and test results. Sometimes such considerations lead to the conclusion that emphases have been wrong and that the curriculum should be changed. In other cases, however, the conclusion may be that the local pattern is quite satisfactory and that no change is warranted. Sometimes the conclusion is that it is better to develop a skill more slowly than implied by the norm in order to reap more noteworthy gains at a later stage in development. Only the local educational authorities, knowledgeable about local objectives and informed about what teaching and learning activities have been occurring, are in a position to make such judgments. And such judgments will be more appropriate if the test results for any particular group in any particular year can be interpreted in the light of similar test results for that same group in past years and for other groups of children of the same ages in past years in the same community. There is no substitute for informed judgment based on accumulated experience in the local setting over a period of time. Unfortunately, there is a tendency to look for simple answers, and comparisons with national norms seem to provide such simple answers.

A second model recommended for use in Title I evaluations calls for the identification of a comparable group of pupils with which a group receiving Title I treatment may be compared. This can be accomplished in a rigorous manner by assigning pupils at random to treatment and comparison groups. But this kind of experimentation is rare in education, and for very good reasons. People take a dim view of having their children used as subjects in experiments. If the new treatment looks as if it will be a good one, everyone wants the new treatment for one's own child. If there's some doubt about the outcome, everybody would just as soon have us continue with the traditional treatment. We sometimes settle for applying the new treatment to those children with parents enthusiastic about the new treatment and the old treatment to those children with parents skeptical of the new treatment—but as any expert in experimental design will tell you, this is no way to conduct an experiment.

I could spend several days going into details regarding the difficulties of making sense out of studies using comparison groups. The conclusion would be that attempts to draw conclusions from such comparisons on the basis of test scores alone and without using professional judgments to make adjustments for the effects of variables assumed to be controlled, but which aren't, generally constitute a misuse of test results.

A third model for Title I evaluation assumes that the performance of those pupils in the schools that are not provided with Title I treatment is adequate as an indicator of what would be a reasonable expectation if no Title I treatment had been provided. But stop and think for a minute. Suppose we have a class of 30 pupils and that on the basis of the pretest we select the lower scoring third for Title I treatment leaving the remaining 20 to be taught by the regular teacher. Does it seem reasonable to suppose that the learning of those 20 will be the same as it would have been if the other ten were included with the group? So far as I can see, there's no way that one can provide special treatment for a subset of a group of pupils without at the same time changing the treatment for the remaining pupils. The regression analyses growing out of applying this third model may look impressive; the underlying assumption is usually not met and so the conclusions are usually invalid.

Many educators, aware of the limitations of scores on standardized tests of achievement--whether the reference norms be national, regional, state, or what have you—have been searching for a more useful approach to the evaluation of educational programs and the monitoring of learning of individual pupils. In recent years many have come to the conclusion that the answer is to be found through the use of criterion-referenced tests. The effort has been encouraged by the development of large pools of test questions by such organizations as the Center for the Study of Evaluation at the University of California at Los Angeles and by some of the national test publishers. The approach seems very attractive. Identify clearly just what it is you want to teach. Prepare or select test questions that measure directly each objective. Decide what proportion of the questions one will consider an acceptable performance for a pupil that has achieved the objective. Arrange the objectives in a logical order and check off each as each pupil meets the objective. Given a reasonable space for posting the checklist, one can at any time tell at a glance just where each pupil is and what needs to be receiving attention next.

Now this attention to purposes and to sequences has much to recommend it. Too often, by trying to teach everything one thinks may prove useful, the school has failed to teach very important elementary knowledge and skills. But to hear some of the advocates talk, one would think that by moving to criterion-referenced tests one can escape all the problems of unreliability and validity one encounters with norm-referenced tests. This simply isn't so. First of all, one needs to be aware of the fact that by concentrating on certain objectives one is inevitably neglecting others. As Professor Robert Pace of UCLA has so clearly told us, "It isn't sufficient to ask, 'Did I reach the objective? One needs rather to ask, What were the consequences?' "[3] This was indelibly engraved on my consciousness many years ago by Dr. William Heard Kilpatrick in a speech at Teachers College, Columbia University, when he told us that as a young teacher in Georgia he was troubled to find that many pupils while learning arithmetic computation were at the same time learning to hate arithmetic and resolving to have nothing to do with it once they could escape from school.

I suppose that if one wishes to make routine calculators of children it doesn't matter much how they go about getting the answers so long as they get the right answers, but my studies with Dr. Harry G. Wheat at West Virginia University some 40 years ago convinced me that if one is interested in developing ability to think mathematically it makes considerable difference how arithmetic is taught and how children go about getting answers to computation problems. An article in a special education section of the *New York Times* of Sunday, April 22[4], indicates that these same concerns are being expressed by mathematic educators today. Criteria stated in terms of a given percentage of right answers to questions of a certain class may be as far from the mark as standards stated in terms of progress along a normative scale. In the first place there's the serious question of what kind of questions we should include. Shall we ask simply that the pupil be able to supply the right answer, or shall we ask also that the pupil be able to demonstrate a sound logical procedure for arriving at the right answer? But even if this question is resolved there remains the question of what shall constitute a satisfactory performance? How many questions shall one include in the criterion-referenced test, and how many of these must be answered correctly to make the performance acceptable? Such research as is available indicates that this is not an easy question to resolve.[5,6]

I find myself uncomfortable talking about criterion-referenced tests in contrast to norm-referenced tests. The fact is that any good achievement test should be a sample from some criterion domain and that what we should be talking about is the breadth or specificity of that domain and about whether the interpretation of the test score is norm-referenced or criterion-referenced. In this sense the examinations I took in elementary school in the 1920's were criterion-referenced so far as interpretation was concerned: the score was always the percentage of possible right answers and the criterion for a pass was 70 percent. Have you ever wondered why we began using norm-referenced interpretations in place of these clearly criterion-referenced tests? One reason certainly was that we came to realize that the score one obtained often depended more on the particular sample of questions chosen by the teacher than on the amount of knowledge one possessed.

Educational practices move in cycles as awareness of the deficiencies of current practices grows; and we often manage to change the words we use to talk about a particular practice when it reappears in a cycle. It's difficult to tell the difference between many of today's criterion-referenced tests of arithmetic and the diagnostic tests of arithmetic that were widely used fifty years ago.

But to return to the problem of concern in this conference. The problem of reliability does not disappear when we substitute a criterion-referenced interpretation for a norm-referenced one. All we do is change the nature of the question we ask about the score. With a norm-referenced interpretation we need to ask, how good is the obtained score as an estimator of the true score of the pupil on the developmental scale used for reporting? With the criterion-referenced interpretation we need to ask, what is the probability that I have classified the pupil correctly? Incorrectly? It makes me very nervous when I see a test that provides three questions for each criterion and which recommends that one interpret the results as indicating that

any pupil who gets two or three questions right has met the criterion while those pupils who obtain zero or one correct answer have not met the criterion. In the first place, I don't find any indication of how it was decided that the criterion should be two questions right and not one or three. In the second place, I see no evidence that anybody has considered the possibility that a person who answers all three questions correctly today might not be able to do so tomorrow—or to answer three new questions on the same content. Such careful analyses as I have seen suggest that it will probably require a minimum of eight questions to provide an acceptable degree of reliability for a classification for any domain and considerably more questions if the domain is other than very simple and specific.[7] I suspect that most teachers will take the results of criterion-referenced interpretations as tentative and subject to change as additional evidence accumulates, but I'm not at all sure that the same wisdom will guide those who use criterion-referenced scores for reaching administrative decisions.

I'm particularly concerned by the nationwide movement to mandated minimum competency testing in the schools, not because I think we should continue to accept what is, rather than to establish reasonable standards, but rather because I think that no single standard can be sound either educationally or politically. I have always thought that it is the task of the school to develop the potential of each child, and when I see the vast differences in the developmental levels of children as they first come to school and as they progress through the school, I am awed by the terrible responsibility that falls on us as educators. We certainly need all the help we can get in learning about the developmental level of the pupils we teach, and both standardized tests and teacher-made tests can help us with this task. If we're using our general tests of achievement appropriately, there's no reason why we can't identify pupils with special learning problems early in their school careers and give them special attention. But we should be charting the progress of *all* pupils, not simply of those exhibiting serious learning deficiencies. It's every bit as serious to fail to challenge a bright pupil to work up to potential as it is to neglect to develop the limited potential of a slow learner.

In this regard I'd like to recount two experiences that throw into perspective the complexity of the problem being attacked by minimum competency testing. The first concerns an account in my local paper some months ago about a truck driver who had registered in the "right to read" program at the local community college. It seems the man had managed to complete the local high school without ever learning to read, and now at last he was going to have this terrible deficiency removed. It turned out that this person had been referred to the university reading clinic while still in the elementary school and that he had been a client of the university reading clinic for six years during which the specialists had tried every approach known for teaching him to read, all without success. His failure to learn simply wasn't because the problem had not been diagnosed. We simply didn't know enough to solve his learning problem.

The second story concerns an individual who was my neighbor years ago in a community far from either Kansas City or Iowa City. I remembered him as a sweet, but not too bright, youngster who babysat for my twin daughters and who was very satisfactory in that capacity. I visited in the community not too long ago and the friend I was visiting asked, "Do you remember Eddie?" (That's not his name, but it will serve for the purposes of this story.)

"Yes," I replied. "What about him?"

"Well," said my friend, "Eddie's the chief gardener at the local factory. He's making an excellent living; he has a wonderful family of six children; the grounds of the factory are the talk of the community; and everybody considers him a solid citizen." I couldn't help thinking, "Wouldn't it have been too bad if they'd had a minimum competency testing program that kept Eddie from graduating from high school so that he wouldn't have been eligible for employment at the local factory?" And a similar thought crossed my mind when I read about the truck driver. Generally, our problem is not that deficiencies haven't been identified; the problem is that in some cases we fail to enlist the effort of the pupil or lack the knowledge to guide the learner effectively.

If funds now being spent on the development and operation of minimum competency testing programs were directed at efforts to understand better how to work effectively with children who are not learning at a normal rate, we would be more likely to solve the problem.

This is not to say that we should continue to certify as competent pupils who for one reason or another have given evidence of failure to learn what they seem to be capable of learning. But the decision regarding whether to pass or fail should be based on a careful clinical assessment, not on a single test score.

So where does this leave us as we think about the *Uses and Abuse of Testing* and about *Tests Relative to Instruction*? First, tests are professional tools with all that is implied by the word *professional*, and we'd better learn to use our tools intelligently if we want to do our job right. In addition we've got to know enough about our tools to resist the oversimplification that we find both inside and outside the profession. Second, tests are indicators to be used as guides to judgment; they are not substitutes for professional judgment. Third, both norm-referenced and criterion-referenced interpretations of test scores are useful and informative. To set the two up as antagonists is to miss the point. Fourth, we have a major task confronting us in helping all people concerned about education to understand what tests can and can't do and to convince legislators, school board members, parents, and the general public that we're not trying to protect vested interests when we insist on resisting attempts to oversimplify test interpretation. And finally, we can't solve our evaluation problems by throwing out the tests. With all their limitations they're one of our most dependable sources of information about what children are learning. As professional educators we need to increase our skill in interpreting test results. As legislators we need to be wary of approving legislation that substitutes rigid regulations about how tests shall be used in the schools for the truly professional use that is often present already. As parents we need to ask more questions about the significance of the test results our children bring home. As citizens we need to cultivate an awareness of how little of the vast richness of what goes on in school can be represented in any set of formal test scores and nip in the bud the temptation to settle for a simple-minded answer to the really complex question, "What are children learning in our schools?"

References

1. Hieronymus, A. N., and Lindquist, E. F. *Manual for Administrators, Supervisors, and Counselors, Iowa Tests of Basic Skills, Form 5 & 6.* Boston: Houghton Mifflin Company, 1974. Pp. 61-63.
2. Thorndike, R. L., and Hagen, E. P., *Measurement and Evaluation in Psychology and Education, Fourth Edition.* New York: John Wiley & Sons, 1977. Pp. 98-100.
3. Pace, R. *Thoughts on Evaluation in Higher Education.* Iowa City, Iowa: American College Testing Program, 1972.
4. Rubin, N. "Beyond the 'New' Math." *New York Times,* Sunday, April 22, 1979, Section 12, pp. 1,30.
5. Glass, Gene V. "Standards and Criteria." *Journal of Educational Measurement,* 1978, *15* (4), 237-261.
6. How to Set Minimum Competency Standards. *How to Evaluate Educational Programs,* 1979, *3* (4), 1-6.
7. Novick, M. R., and Lewis, C. "Prescribing Test Length for Criterion-Referenced Measurement." In C. W. Harris, M. C. Alkin, and W. J. Popham (Eds.) *Problems in Criterion-Referenced Measurement* (CSE Monograph Series in Evaluation: 3). Los Angeles: Center for the Study of Evaluation, UCLA, 1974. Pp. 139-158.

Dr. Coffman is Lindquist Professor of Education and Director of the Iowa Testing Programs at The University of Iowa. He has conducted research and written extensively in the areas of achievement and aptitude testing and the evaluation of educational programs. Before joining The University of Iowa, Dr. Coffman taught in both elementary and secondary schools, held major administrative appointments at Educational Testing Service, including Director of Test Development and Director of Research and Development, and was a visiting lecturer at several major universities.

THE EXECUTIVE REVIEW
Published eight times a year by

The University of Iowa
210 Lindquist Center
Iowa City, Iowa 52242

Institute for School Executives

William C. Bozeman, *Editor*
George A. Chambers, *Director*